WASHINGTON STATE

GEOSPATIAL DATA INITIATIVE

MANAGING GEOSPATIAL DATA ACROSS MULTIPLE JURISDICTIONS AND DISCIPLINES

DISCUSSION DRAFT   April 4, 1995   DISCUSSION DRAFT

Geographic Information Council

Local Government Workgroup

EXECUTIVE OVERVIEW

Geographic information systems are a powerful tool for supporting the business activities of an organization. Their ability to store, manipulate, and display mapping information electronically, and often on short notice, allows managers to make more informed decisions. Their presence has taken management to new heights of efficiency and effectiveness.

However, there is a down side to geographic information systems. Data are currently collected and stored independently by different organizations for specific uses. These data are often redundant and inconsistent between organizations, making it very difficult to combine the data and produce meaningful results. If this situation continues, the data in geographic information systems will become as disparate as the data in traditional information systems, and will be just as difficult to correct.

The data management initiative, concepts, and strategies for controlling the data in geographic information systems are presented in this paper. The presentation begins by describing the current situation of increasing quantities of disparate data, and the problems that will arise if the data in geographic information systems are not controlled. The State's data initiative is explained and a geospatial data initiative is presented.

The concepts for managing the data in geographic information systems are outlined, beginning with a definition of terms commonly used with geographic information systems. The four-tier, multiple-extent, multiple-chronology, and multiple-resolution concepts are presented with examples. The combination of these new concepts to form a three-dimensional geospatial data aggregation scheme are described. The georeferencing between tabular data and geospatial objects is also described.

Based on the pending problems with geospatial data, the geospatial data initiative, and the concepts for managing the data in geographic information systems, a set of strategies is presented for controlling geospatial data. These strategies are divided into three groups for data architecture, data awareness, and data accuracy. The data architecture strategies include maintaining a common data architecture, establishing a hierarchy of data themes and sub-themes, identifying framework data layers, integrating geospatial data, and designating data responsibilities. The data awareness strategies include identifying metadata, implementing a data clearinghouse for those metadata, conducting a general geospatial data survey, conducting a detailed geospatial data inventory, and identifying topics and keywords for the data clearinghouse. The data accuracy strategies include designating control datums, establishing geospatial data certification criteria, establishing geospatial data enhancement coordination, establishing geospatial data aggregation criteria, and establishing geospatial data version criteria.

The implementation of these data strategies ensures that the data in geographic information systems are developed consistently in the State, adequately represent the real world, can be readily shared between organizations, and add value to the decision making process.

CONTENTS

CURRENT SITUATION

The situation in most organizations today is massive quantities of disparate data, continued production of disparate data at an alarming rate, and increased demand for integrated data to meet business needs. This situation has created a data paradox for those organizations. Information systems and databases are developed faster with client-friendly development tools to meet local business needs. However, that rapid development is contributing to the increasing quantities of disparate data which can not be easily integrated for more global business needs.

A variety of tools are currently available to electronically connect different databases and access their data. However, these tools only provide electronic access to the data. They do not provide any understanding of the data in those databases, or any method to convert the data in those databases to a common form once the data are understood. Many organizations are not aware of this distinction and are assuming that disparate data can be easily integrated with tools that are currently available.

The present disparate data situation exists largely with tabular data developed over the last several decades. However, the disparate data situation is just beginning with geospatial data. The high interest in geospatial data, the increased capability of geographic information systems, and the reduced cost of geographic information systems and hardware dramatically increases the rate of geospatial data capture and storage. The extension of geospatial data from environmental and Earth related interests into demographic interests further increases the rate of geospatial data capture and storage.

The impending problem with geospatial data is rapidly increasing quantities of disparate data that can not be easily integrated, the same as most organizations are already facing with their tabular data. If geospatial data are not controlled, they will become as disparate as tabular data within a few years, which greatly increases the difficulty of integrating data to meet business needs. In addition, the task of georeferencing disparate tabular and disparate geospatial data become extremely difficult.

Another trend currently evolving is the use of data warehouses to store large quantities of historical tabular data for the analysis of trends and projections. One major problem with implementing data warehouses is the huge quantities of disparate tabular data and the transformation required to put those disparate data in a common form. However, this is only the beginning of the problem. Eventually data warehouses will store historical geospatial data for analysis of trends and projections the same as tabular data. If geospatial data become as disparate as tabular data, the problems mentioned above increase substantially.

DATA INITIATIVE

In 1988 an initiative was established by the Policy and Planning Division of the Department of Information Services to *'Define data in a common context so they could be readily shared.'* This initiative was the first step toward gaining control of the State's disparate data by limiting the production of disparate data and integrating existing disparate data. The initiative applies to all data, including tabular data, geospatial data, and historical data.

The original scope of this initiative was core State agencies. However, with an increasing awareness of the problem, the scope expanded to include State agencies, city and county government, Indian tribes, public utilities, and Federal agencies. The reason for the expanded scope was the urgent need to integrate data across both multiple jurisdictions and multiple disciplines. In other words, the need for integrated data to meet business needs was not limited to the data in core State agencies.

On March 23, 1989, the Information Services Board adopted a policy statement regarding managing data as a resource of the State.

> *Data are to be managed as a resource and asset of the State and must be consistently named and defined to facilitate data sharing across business activities and organizational boundaries.*
>
>> *The Information Services Board will establish prospective State-level standards for management of the State's data resource.*
>>
>> *The Department of Information Services will identify common, high-impact areas of the data resource and will facilitate efforts to consistently name, define, and document the data in those areas.*
>>
>> *State agencies will assure that any new data, new applications, or major application and database upgrades conform to the data resource standards defined by the Department of Information Services.*
>>
>> *Each State agency will maintain current, accurate documentation of the data used by that agency and assure that documentation is readily available.*

This policy applies to both tabular and non-tabular data, including geospatial data. Non-tabular data are as much a part of the State's data resource as tabular data, and must be managed with the same intensity as tabular data. To emphasize the importance of geospatial data, a separate geospatial data initiative was developed.

> *Gain control and increase the understanding of geospatial data before they become as disparate as tabular data, and integrate geospatial data with tabular data to meet business needs and support business strategies.*

DATA MANAGEMENT CONCEPTS

The geospatial data initiative is met by using a data management approach based on the integration of four concepts: the four-tier concept, the multiple-extent concept, the multiple-chronology concept, and the multiple-scale concept.

Basic Definitions

An explanation of the four concepts supporting the geospatial data initiative begins with a definition of basic terms.

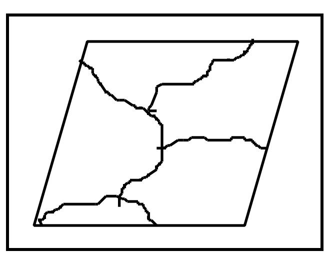
Geospatial is a term derived from *geo* meaning Earth and *spatium* meaning of, relating to, involving, or having the nature of space. Literally it means spatial locations on the Earth.

A *geospatial object* is any natural feature, constructed feature, or boundary on, above, or below the Earth's surface. For example, a stream, a school, a county, and an aquifer are geospatial objects.

Geospatial data are any non-tabular data that represent the geographic location and identifying characteristics of a geospatial object. For example, the location of the beginning and ending of the stream and the stream name, the location and name of a school, and the boundaries and name of a county are geospatial data.

Tabular data are any data that are data stored and displayed in tabular form. Most traditional data in database management systems are tabular data.

A *data layer* is any separate and distinct set of related tabular or non-tabular data that is maintained in a geographic information system. It represents a particular topic of interest in the real world. For example, a data layer might be defined for the county roads in a county, as shown in the diagram below.

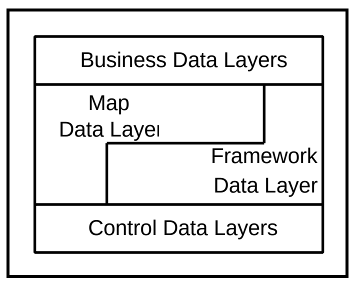


A *data layer extent* is the outer boundary, or limits, of a data layer. For example, the data layer extent for the county roads maintained in a county would be that county's boundary.

The *data layer coverage* is the portion of a data layer extent for which data are available. For example, if the county road data layer included only roads in rural areas, the coverage would be rural areas of the county.

The *extent of interest* is the outer boundary of an area that is of interest to an organization for some business purpose. It may be part of a data layer extent, a complete data layer extent, or multiple data layer extents. For example, if there were interest in all county roads in Western Washington, the extent of interest would be all the counties in Western Washington.

A *control data layer* is a ubiquitous data layer that provides the horizontal and vertical (x, y, and z) control necessary for aligning other data layers and accurately placing geospatial objects on the Earth. It is necessary to keep all data layers properly aligned so they can be aggregated. For example, the Public Land Survey and State Plane Coordinates could be control data layers. Control data layers are the foundation for other data layers, as shown in the diagram below.



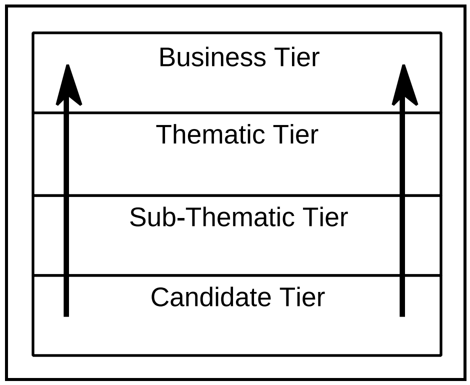
A *framework data layer* is a ubiquitous data layer that is basic to many other map data layers. A framework data layer is based on a control data layer, as shown in the diagram above. For example, elevations, such as a digital elevation models, elevation contours, transportation routes, land parcels, and surface water bodies could be framework data layers.

A *map data layer* is any data layer other than a control data layer or framework data layer that is of interest to an organization. It has a more specific, limited interest than control or framework data layers. A map data layer can be based on either a control data layer or a framework data layer, as shown in the diagram above. For example, population, crime, disease, geologic hazard areas, and vegetation type are map data layers.

A *business data layer* is any data layer resulting from a combination of framework and/or map data layers to meet specific business needs, as shown in the diagram above.

Geospatial Data Tiers

Geospatial data are grouped into tiers representing levels of data aggregation. The *four-tier concept* supports the geospatial data initiative by defining four tiers of data layer aggregation, as shown in the diagram below. Each geospatial data tier contains one or more data layers.



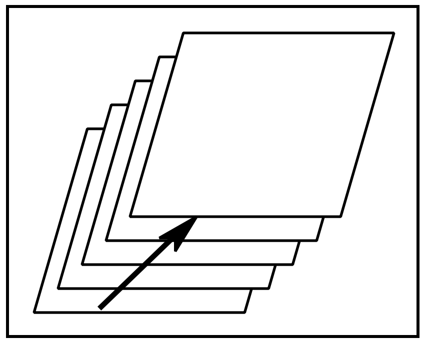
The *Business Tier* consists of data layers that are a combination of data layers from the thematic and/or sub-thematic tiers to meet business needs. There are relatively few thematic and sub-thematic data layers, but there can be many combinations of those data layers to meet business needs. The data layers in the business tier may be permanent or temporary depending on the continued business needs. For example, data layers for roads, rivers, and bridges may be combined to meet a specific business need.

The *Thematic Tier* consists of data layers that represent a major data theme, such as transportation, land use, surface water, or waste water infrastructure. The data layers in the thematic tier may be a control data layer, a framework data layer, or a map data layer. Data layers in the thematic tier are usually permanent, but may be temporary or non-existent depending on the use and importance of data layers in the sub-thematic tier. For example, data themes may be designated for transportation, surface water bodies, and vegetation.

The *Sub-Thematic Tier* consists of data layers that contribute to a major data theme. These sub-thematic data layers collectively contribute to a fully enhanced thematic data layer. There are usually one or more permanent data layers in a sub-thematic tier, but there may be none if there is a single layer in the thematic tier. For example sub-thematic data layers might be developed for waste water facilities, interceptor sewers, storm water drains, gravity sewers, and storm water outfalls that contribute to the waste water infrastructure data theme.

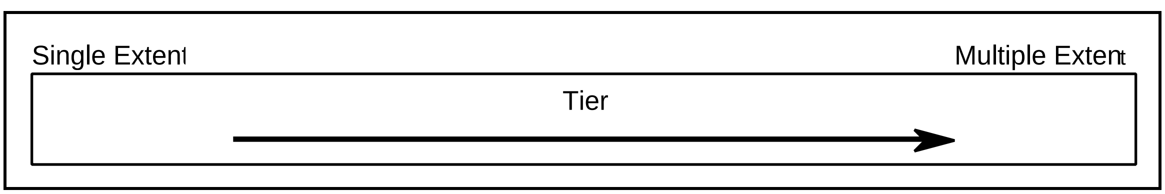
The *Candidate Tier* consists of data layers that are candidates for building sub-thematic data layers. They are temporary data layers for entering and adjusting data, deciding what data to combine into a sub-thematic data layer, or determining how to combine data to form a sub-thematic data layer. For example, candidate data layers might be defined for pumping stations, treatment plants, interceptor sewers, settling ponds, manholes, outfalls, storm water drains, storm water outfalls, and so on, for the waste water infrastructure.

Data layers can be aggregated vertically within or between tiers to form a more enhanced data layer, as shown in the diagram below. *Vertical data layer aggregation* is the combination of two or more data layers to form a more enhanced data layer. For example, the waste water facilities, interceptor sewers, storm water drains, gravity sewers, and storm water outfalls data layers could be combined to form the waste water infrastructure data layer.

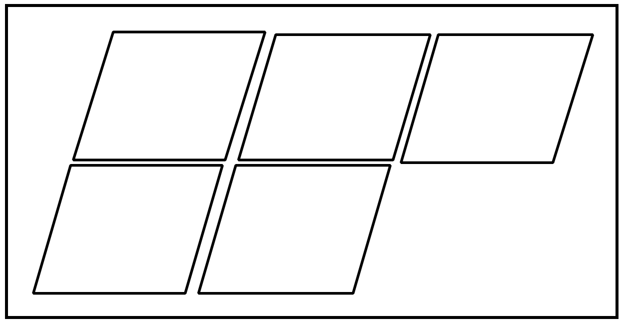


Geospatial Data Extents

Geospatial data occur in single extents and a combination of extents. The *multiple-extent concept* supports the geospatial data initiative by defining a continuum of data layer extents from an individual extent maintained by an organization to the aggregation of multiple extents from many organizations to meet a specific extent of interest, as shown in the diagram below.

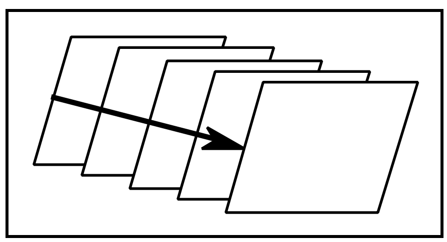


Data layers can be aggregated horizontally to form a larger data layer extent. *Horizontal data layer aggregation* is the edge connection of two or more data layer extents, usually from different organizations, for the same data layer to provide an expanded data layer for a specific extent of interest, as shown in the diagram below. For example, the county road data layers from several counties could be combined to show the county roads for Western Washington.



Geospatial Data History

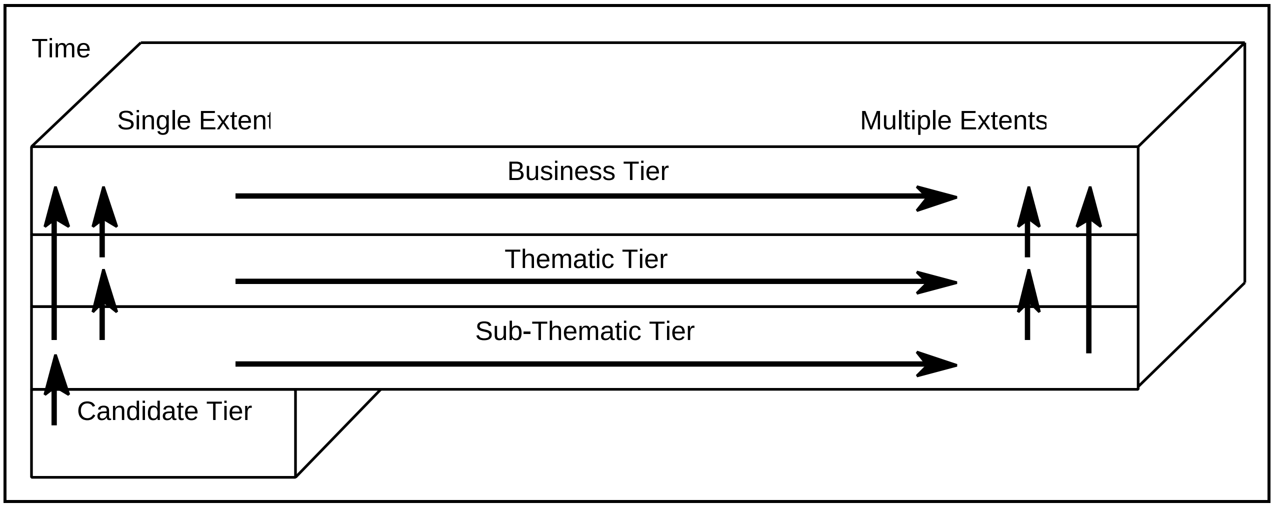
Geospatial data represent the real world at a point in time. The *multiple-chronology concept* supports the geospatial data initiative by defining a time series of data layer extents, as shown in the diagram below. This concept is important both for analyzing trends and making projections, and for the storage of geospatial data in data warehouses for more automated analysis of trends and projections. Any data layer may have multiple time periods.



Data layers can be aggregated chronologically to provide a history of changes. *Chronological data layer aggregation* is the display of successive time periods for a specific data layer extent. For example, the spread and recession of surface water pollution could be tracked over time and compared to pollution abatement efforts.

Combined Concepts

The four-tier concept, multiple-extent concept, and multiple-chronology concept are combined to support the geospatial data initiative, as shown in the diagram below. This combined concept provides the base for coordination and cooperation among organizations and across disciplines to meet both local needs and combined needs.

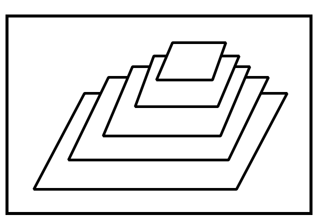


The combined concept consists of a three-dimensional aggregation of geospatial data to meet business needs. Vertical aggregation combines data layers to form a more enhanced data layer. Vertical aggregation occurs between any tiers, as shown by the vertical arrows on the diagram. Horizontal aggregation connects data layer extents for the same data layer to provide a larger extent. Horizontal aggregation occurs in any of the top three tiers, as shown by the horizontal arrows on the diagram. Chronological aggregation combines time periods of data layers to provide a history of changes. Chronological aggregation occurs in any of the top three tiers.

The three-dimensional aggregation of data layers forms a network of aggregations to meet business needs. The difficulty with this network of aggregations is a loss of control over geospatial data. Different organizations perform different aggregations, on different versions of data, using different algorithms to meet different business needs. Many of these aggregations are not completely documented, resulting in a loss of understanding about the aggregated data layers.

Geospatial Data Generalization

The *multiple-scale concept* supporting the geospatial data strategy represents levels of generalization of geospatial data, as shown in the diagram below. Geospatial data are captured and stored at a specific scale, such as 1:12,000. Those data can be generalized to a smaller scale, such as 1:25,000, 1:100,00, or 1:250,000. There may be several levels of generalization for any geospatial data tier.



Data layer generalization is the process of reducing the scale of a data layer extent from to the scale at which the data were captured by a formal algorithm. For example, the scale for surface water bodies may be reduced from 1:12,000 to 1:100,000 to gain a broader perspective for surface water bodies. However, with the scale reduction there is a loss of detail and many smaller surface water bodies may not appear. The scale can not be increased from the scale at which the data were captured because it presents a false accuracy.

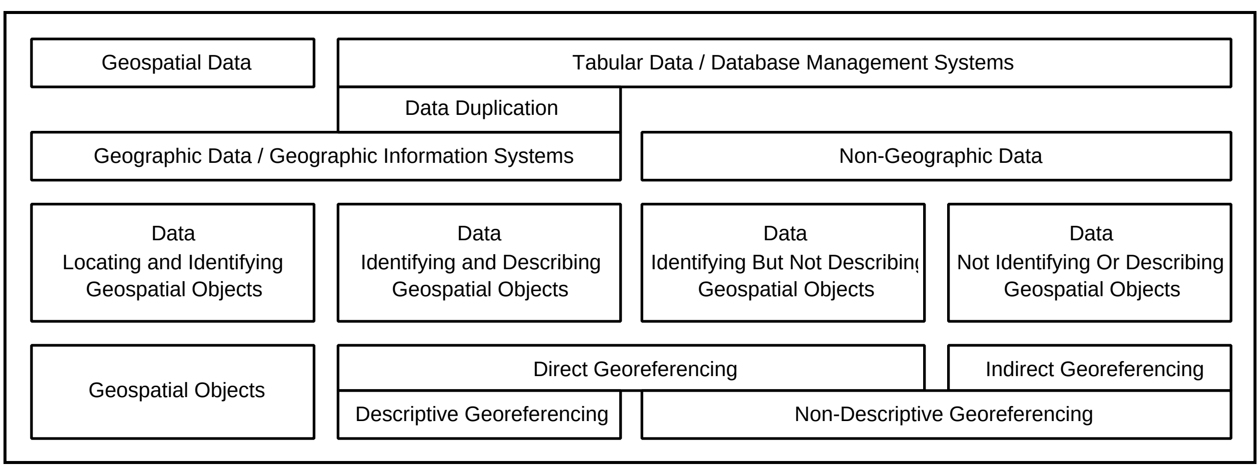
DATA GEOREFERENCING

With the increasing popularity of geospatial data, there is an increasing interest in connecting the large quantities of existing tabular data to geospatial data. The process of georeferencing tabular data to geospatial objects provides additional meaning to relatively meaningless boundaries. For example, the boundaries of a school district are relatively meaningless without some data to display within those boundaries. If tabular data were analyzed, such as average grades or race, and displayed by school district, the boundaries would have increased meaning.

The ability to create a reference between tabular and geospatial data, particularly with tabular demographic data, will make massive quantities of data available for spatial display. For example, on-line analytical processing in data warehouses could produce population distributions for elementary school students that could be displayed by school district. The trend in population distributions over time could also be displayed spatially. Therefore, accurate georeferencing is mandatory for connecting tabular data to geospatial objects.

Definitions

The overall concept for georeferencing is shown in the diagram below. There are four groupings of data in the center of the diagram representing how tabular and non-tabular data are related to geospatial objects. The first grouping contains geospatial data that locate and identify geospatial objects. It does not contain any data that describes geospatial objects. The second grouping contains tabular data that identify and describe geospatial objects. It does not contain any geospatial data that locate geospatial objects. The third grouping contains tabular data that identify but do not describe geospatial objects. The fourth grouping contains tabular data that do not identify or describe geospatial objects.



The three rows at the top of the diagram shows the relation between geospatial data and tabular data, and between geographic data and non-geographic data. The top row shows geospatial data and tabular data, and the third row shows geographic data and non-geographic data. Tabular data are typically stored in database management systems and geographic data are typically stored in geographic information systems. The second row shows the potential duplication of tabular data between database management systems and geographic information systems.

> *Geographic* is a term derived from *geo* meaning Earth and *graphein* meaning to write or describe. Literally it means describing the Earth.

> *Geographic data* are geospatial or tabular data that locate, identify, or describe geospatial objects. They may be stored in either database management systems or geographic information systems. There is a potential for duplication of tabular data between database management systems and geographic information systems. These duplicate data must be properly maintained to prevent any discrepancy.

> *Non-geographic data* are tabular data that may or may not identify geospatial objects, but do not describe those geospatial objects. They are stored in database management systems.

The bottom of the diagram shows the different types of georeferencing that can occur. The boxes on the right represent the different types of georeferencing that can occur between tabular and geospatial data.

> *Georeferencing* is the connection between a set of tabular data and related geospatial data through common identifying characteristics so the tabular data can be displayed spatially. The term is derived from *geo* meaning Earth, and *referre* meaning to carry back.

> *Geocoding* is the process of designating unique identifiers for geospatial objects and maintaining those unique identifiers in both the geospatial data and related tabular data so that georeferencing is possible. Georeferencing can not occur without the establishment of unique identifiers.

> *Descriptive georeferencing* occurs between a set of tabular data that identify and describe a geospatial object, and geospatial data that locate and identify that geospatial object. The length, upper elevation, lower elevation, maximum flow, and minimum flow of a stream; the date of construction, capacity, and value of a school; and the size, land area, and population of a county are tabular data describing a geospatial object that can be georeferenced to that object.

> *Non-descriptive georeferencing* occurs between a set of tabular data that either identify but do not describe geospatial objects or do not identify or describe a geosaptial object, and geospatial data that locate and identify that geospatial object. Data about the analysis of water samples from a stream, data about students in a school, and data about customers living in a county are tabular data that do not describe a geospatial object, but can be georeferenced to that object.

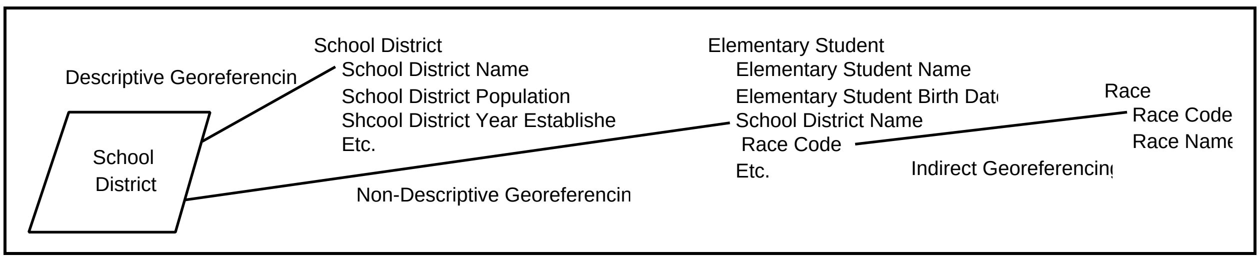
> *Direct georeferencing* occurs between a set of tabular data that either identify and describe a geospatial object or identify and do not describe a geospatial object, and

geospatial data that locate and identify that geospatial objects. The georeferencing occurs directly through the geospatial object identifiers.

Indirect georeferencing occurs between a set of tabular data that do not identify or describe a geospatial object, and geospatial data that locate and identify that geospatial object. Indirect georeferencing occurs indirectly through another set of tabular data that carry the geospatial object identifier. Data about the method of water sample analysis, data about the race of school children, and data about the population in a county are tabular data that do not describe a geospatial object and can only be georeferenced to that object through another tabular data set.

Georeferencing Examples

The diagram below shows examples of descriptive and non-descriptive georeferencing, and direct and indirect georeferencing.



Descriptive georeferencing is shown on the left of the diagram for school districts. The geospatial objects are School Districts with School District Name as the unique identifier. The tabular data describe each School District and contain the unique identifier. A descriptive georeference occurs between the tabular and geospatial data using School District Name.

Non-descriptive georeferencing is shown in the center of the diagram for elementary students. The geospatial object is still School District with School District Name as the unique identifier. The tabular data for Elementary Students do not describe the School District, but it contains the unique identifier of School District Name. A non-descriptive georeference occurs between the tabular and geospatial data using School District Name.

These two examples represent direct georeferencing because the unique identifier for a School District, School District Name, is stored in both the School District and Elementary Student tabular data sets. The tabular data can be directly referenced to the geospatial objects representing School Districts.

Indirect georeferencing is shown on the right of the diagram for the Race of Elementary Students. Race does not contain the identifier for School District, so a georeference can not be made directly between Race and School District. It must be made indirectly through Elementary Student which does contain the unique identifier School District Name. In other words, the Race is determined for each Elementary Student which is summarized and displayed by School District.

DATA STRATEGIES

Geospatial data strategies are necessary to implement the geospatial data initiative. These strategies assure the consistent development of geospatial data and the proper integration of geospatial and tabular data to meet business needs. The geospatial data strategies are grouped by data architecture strategies, data awareness strategies, and data accuracy strategies. One or more tasks are defined for each geospatial strategy.

Data Architecture Strategies

Geospatial data development must be carried out within a consistent architecture that encompasses all data in the State's data resource. The following data architecture strategies ensure that geospatial data for the State are developed and maintained in a consistent manner.

Maintain a Common Data Architecture

A common data architecture is being developed for the State to meet the initiative of defining all data in a common context so they can be readily shared.

> Include the development of geospatial data within the State's common data architecture to ensure commonality among jurisdictions and across disciplines. The Geographic Information Council, administered by the Department of Information Services, should be the lead organization for the common data architecture.

Establish Data Themes and Sub-Themes

The definition of thematic and sub-thematic data layers benefits the State in three ways. First, there will be minimal design effort because thematic and sub-thematic templates provide the data layer design. Second, the data layer extents will easily edge-match between adjoining jurisdictions without major adjustment. Third, data layers will easily aggregate vertically to meet business needs without alteration or conversion.

> Develop a hierarchy of thematic and sub-thematic data layers that include current needs and future needs of organizations within the State. The hierarchy should be consistent with the common data architecture being developed for the State. An inter-disciplinary team of domain experts should be established to define a hierarchy that meets all needs.
>
> Designate common unique identifiers for geospatial objects in each data layer. These common unique identifiers should be used by all organizations to designate geospatial objects and to create georeferences between tabular data and geospatial objects.

Provide geospatial thematic templates for developing data layers within each organizations. Templates should be defined for all control, framework, and map data layers. Initial emphasis should be placed on providing templates for the control and framework data layers because they provide the foundation for aggregating geospatial data. Templates for map data layers can be provided after the control and framework templates are available.

Identify Framework Data Layers

Framework data layers provide the foundation for developing map data layers and aggregating data layers to meet business needs. The definition and development of common framework data layers that can be used by all organizations in the State benefit the State at large by preventing development of multiple versions of framework data layers, by ensuring consistent updates for the framework data layers, and by minimizing the effort of organizations using the framework data layers.

Identify the framework data layers needed for the State within the thematic / sub-thematic hierarchy described above. All the details for each framework data layer should be defined, including source data, scale, resolution, precision, etc. The organization responsible for maintaining each framework data layer should also be defined. An inter-disciplinary team of domain experts should be established to define the framework data layers

Establish a priority and schedule for developing and implementing the framework data layers.

Integrate Geospatial Data

Considerable geospatial data already exist in many organizations. However, the existing data are not necessarily consistent across those organizations. To maximize the use of existing geospatial data, the existing data layers should be aligned to the thematic / sub-thematic hierarchy defined above. Even though the alignment could result in an additional effort to some organizations, the result would be geospatial data that are more readily shared among organizations.

Encourage organizations to align existing geospatial data layers to the thematic / sub-thematic hierarchy and to development new geospatial data layers within this hierarchy. Each organization maintaining geospatial data should be responsible for identifying the prominent data layers they maintain and should align those data layers with the thematic / sub-thematic hierarchy.

Designate Data Responsibilities

Geospatial data are captured by many organizations for many different uses. In many situations, the capture and maintenance of geospatial data is redundant and inconsistent, and uses valuable resources unnecessarily. The designation of primary organizations responsible for the capture and maintenance of geospatial data reduces the effort in collecting those data, and provides more consistent data to other organizations.

> Identify organizations that have primary responsible for the maintenance of data layers in the thematic / sub-thematic hierarchy. Several organizations may need to be designated if there are several extents for a data layer. These organizations should be the primary source for geospatial data. A current list of designated organizations should be maintained and made readily available to anyone desiring to know where data can be obtained.

> Encourage other organizations to obtain geospatial data from the primary organization designated for maintenance of those geospatial data rather than redundantly capturing or updating those data.

Data Awareness Strategies

Organizations must be aware of the data that exist or are planned and thoroughly understand those data so they can be used effectively and efficiently to meet business information needs. The following strategies ensure that organizations are aware of existing and planned data before they proceed with any geospatial data development efforts.

Identify Metadata

Metadata are data about the data, or simply data that help people understand the data. They are data needed to easily identify, readily obtain, and effectively use data to meet business needs. As the State's data resource becomes larger and more complex, the need for complete, current, accurate, and readily available metadata is mandatory. Without proper metadata, the State's data resource can not be effectively or efficiently used.

> Define the set of metadata required to document the State's geospatial data and use those metadata to develop and maintain a data clearinghouse. Existing standards should be cross referenced to the State's common data architecture and used to identify all possible metadata of interest to the State. An inter-disciplinary team of domain experts should be formed to select the metadata necessary to document the State's geospatial data.

Encourage organizations using geospatial data, particularly the primary organization responsible for maintaining data layers, to use the metadata to document all data layers in the State.

Implement a Data Clearinghouse

Data awareness can only be achieved if there is some mechanism to make metadata available to a wide variety of organizations scattered throughout the State. The development and distribution of a data clearinghouse helps organizations become aware of data that are currently available and plans for the development of new data. The data clearinghouse should include metadata about all the State's data. It is inappropriate to develop a data clearinghouse for only one segment of the State's data resource because it could compromise support for business needs.

Design, develop, and implement a data clearinghouse for all the State's data, including geospatial data. The data clearinghouse should include the metadata defined above, and should show organizations and organization contacts for the data layers, projects, and documents. The Department of Information Services should develop and maintain the data clearinghouse.

Establish a priority for populating the data clearinghouse based on prominent business needs within the State. A representation of all organizations and all disciplines in the State should be included in the evaluation of priorities for the data clearinghouse.

Conduct Geospatial Data Survey

A data survey is a high-level search for any data that exist in the State and for projects that are currently in progress or planned that pertain to data capture and maintenance. A data survey includes both automated data, such as databases or data layers, and non-automated data, such as maps or documents. Conducting a data survey, and placing the results in a data clearinghouse increases the data awareness and prevents duplication of effort.

Identify a subset of the metadata defined above for use in the data survey. The metadata subset should include any data necessary for identifying and documenting the existence of major geospatial data sets within the State. It is impractical to capture the full set of metadata during a high-level survey. An inter-disciplinary team of domain experts should be formed to select the metadata used in the data survey.

Conduct a survey of geospatial data that currently exist in the State, and any current or planned projects that pertain to the collection, maintenance, and use of geospatial data in the State. The results of the survey should be placed in the data clearinghouse. The Geographic Information Council should lead the survey.

Encourage organizations maintaining geospatial data or planning geospatial data projects in the State to make entries into the data clearinghouse. Organizations should also be encouraged to consult the data clearinghouse before proceeding with a project to determine if data already exist or if other organizations are planning similar projects.

Conduct Geospatial Data Inventory

A data inventory is a detailed documentation of the specific data that exist in a data set. A data inventory is usually conducted after a data survey, and is usually done for prominent data sets identified in a data survey. The results of a data inventory are placed in the data clearinghouse so organizations are aware of the detailed data that are available in prominent data sets.

Designate prominent data sets among those identified in the data survey. An inter-disciplinary team of domain experts should be established to define the criteria for a prominent data set and to designate the prominent data sets.

Conduct a data inventory of the prominent data sets to determine the specific data in each of those data sets. The Geographic Information Council should lead the data inventory.

Cross reference the data to the State's common data architecture so the content and meaning of each data item can be determined within a common context. The Department of Information Services should lead the data cross referencing.

Enter the results of the data inventory and cross referencing into the data clearinghouse so they are available to any organization interested in the details of prominent data sets.

Identify Topics and Keywords

A data clearinghouse is only useful if there is an easy, comprehensive, mechanism to enter and search that clearinghouse. The establishment of a set of business keywords and core topics provides ready access to the data clearinghouse and helps clients locate data sets that could meet their business needs.

Develop a set of core topics and related business keywords for indexing items in the data clearinghouse. The core topics should be subjects for directly accessing the clearinghouse data. The business keywords should include any business terms clients might use to access the data clearinghouse and should be indexed to the topics. An inter-disciplinary team of domain experts should be established to develop the core topics and business keywords.

Define the geographic areas desired for indexing items in the data clearinghouse. The geographic areas should include cities, counties, quadrangles, townships, water resource inventory areas, hydrologic units, and any other prominent geographic areas necessary to access the clearinghouse data.

Data Accuracy Strategies

The accuracy of geospatial data is critical for effective business support. Often the accuracy of data is not known and may not be at the desired level. The current major accuracy problems are documenting the accuracy of existing data and planning the altering of the current accuracy to that desired. The following strategies ensure that the current data accuracy is known and documented, and a base is established for adjusting data accuracy to the level desired.

Designate Control Datums

Designation of consistent horizontal and vertical control datums is mandatory for the accurate placement of geospatial objects on the surface of the Earth. If different control datums are used, geospatial objects are inconsistently placed on the Earth's surface and it is difficult to aggregate data with meaningful results. The establishment of consistent control datums assures that geospatial data can be accurately aggregated and placed on the Earth's surface.

Establish a preferred horizontal datum for geospatial data in the State, such as NAD 83(91).

Establish a preferred vertical datum for geospatial data in the State, such as NAVD 88.

Encourage all organizations to capture geospatial data according to the preferred horizontal datum, or to convert their data to the preferred horizontal datum.

Encourage all organizations to capture geospatial data according to the preferred vertical datum, to convert their data to the preferred datum, or to provide offsets from the preferred vertical datum.

Establish Geospatial Data Certification Criteria

Geospatial data should conform to specific criteria so they can be effectively used to support an organization's business needs. Establishment of geospatial data certification criteria, and alignment of geospatial data conform to those criteria, ensures that geospatial data are the highest quality possible and will meet an organization's business needs.

Establish certification criteria for geospatial data in the State. These certification criteria should specify the preferred level of accuracy for all geospatial data, how well the data layers align with the thematic / sub-thematic hierarchy, and how well they align with the State's common data architecture. The geospatial data certification criteria should meet or exceed the existing Federal requirements for geospatial data. An inter-disciplinary team of domain experts should be established to prepare these geospatial data certification criteria.

Encourage all organizations collecting or maintaining geospatial data about the State to follow these certification criteria, and to stipulate that their data do meet the certification criteria in any documentation about their data.

Establish Geospatial Data Enhancement Coordination

Many geospatial data extents are maintained for the same data layer by different organizations. Each organization often enhances their specific extent independent of the other organizations. This independent enhancement, though good for the organization making the enhancement, causes a discrepancy when data layers are aggregated. The establishment of procedures for consistently enhancing multi-extent data layers, and making those enhancement available to other organizations, assures the effective and efficient use of geospatial data.

Develop a procedure for coordinating the enhancement of multi-extent geospatial data maintained by many different organizations. The procedure should explain in detail how the enhancement of multi-extent data layers should be coordinated and how the information about new enhancements is disseminated to other organizations.

Designate the data layers in the thematic / sub-thematic hierarchy that are multi-extent data layers maintained by different organizations. These data layers should follow the multi-extent coordination procedures. If appropriate, a central repository that contains all current enhancements should be designated for each multi-extent data layer.

Encourage all organizations that enhance multi-extent data to follow these coordination procedures, and all organizations that use multi-extent data to go to the primary source for the latest enhancements.

Establish Geospatial Data Aggregation Criteria

Geospatial data can be aggregated horizontally, vertically, and chronologically to provide more detail. However, the pathway of aggregation and the algorithms used for aggregation can vary widely and are often undocumented, resulting in uncertainty about what an aggregated data layer represents. The establishment of geospatial data aggregation criteria assures organizations that the most efficient pathway for aggregation was used and algorithms for data aggregation are documented.

Develop criteria for aggregating geospatial data horizontally, vertically, and chronologically. These criteria should include determining the most efficient pathway for aggregation and the method of documenting the process used for aggregation. An inter-disciplinary team of domain experts should be established to develop these criteria to assure that they are suitable for all disciplines.

Encourage all organizations aggregating geospatial data to follow these aggregation criteria in the development and documentation of aggregated data layers.

Establish Geospatial Data Version Criteria

Geospatial data are frequently enhanced as the real world they represent changes. These enhancements are performed at different times by different organizations resulting in multiple versions of data for the same data layer. Although these enhancements occur at different frequencies, the documentation of each version of data is useful when data layers are aggregated. The data version documentation assures that people know what time frame the data layer represents.

Develop the criteria for designating versions of geospatial data. These criteria should include the notation for identifying separate versions of data and the method of documenting the time frame those data represent. An inter-disciplinary team of domain experts should be established to develop these criteria to assure that they are suitable for all disciplines.

Encourage all organizations enhancing or using geospatial data to properly designate the data version based on the version criteria, and to include the data version designations in all geospatial data products.

CONCLUSION

The proper management of geospatial data is important for the effective and efficient use of those data, and for the optimum utilization of resources. The time, budget, and personnel constraints that many organizations face today do not allow any leeway for developing data redundantly and inconsistently, or for spending unnecessary resources to collect and maintain data or share data. All resources must be used effectively to support the business activities of an organizations.

The proper management of geospatial data is assured with the establishment of a geospatial data initiative, the definition of data management concepts to support that initiative, and the implementation of strategies to carry out the initiative. The current situation of rapidly increasing quantities of disparate data, and their associated problems, will only get worse unless some method of control over the data resource is implemented. The only way to stop the disparity and gain control is through the implementation of an integrated set of strategies to properly manage data.

Implementation of the fifteen strategies outlined in this paper supports a theme of thinking globally and acting locally. Organizations need to think in terms of a global data architecture for all data and develop local data to meet their specific needs within that architecture. This approach is opposite of the traditional approach of thinking locally and producing global junk. Thinking globally and acting locally provides the best data possible, allows those data to be readily shared across diverse organizations and disciplines, and maximizes the utilization of limited resources.

Thinking globally and acting locally leverages each organizations investment in their data resource. It avoids redundant data collection and unnecessary gaps in the data needed to support business activities. It resolves the paradox of having a high demand for current, accurate data, often on short notice, and not having the data resource available to support that demand. It supports an organization in its effort to be successful in an increasingly dynamic business environment.

For example, an employee's address city name and zip code can be used to spatially display data about employees, as shown in the diagram below. Similarly, the school district number and watershed identifier can be used to spatially display descriptive data about school districts and watersheds.